A MULTIDIMENSIONAL SKETCHING INTERFACE FOR CORPUS BASED CONCATENATIVE SYNTHESIS

Augoustinos Tsiros
Edinburgh Napier University, Centre for Interaction Design,
Edinburgh, EH10 5DT
a.tsiros@napier.ac.uk

ABSTRACT

This paper presents *Morpheme,* a multidimensional interface that allows real-time control of concatenative synthesis through the act of sketching on a digital canvas. *Morpheme* extracts textural, spatial and volumetric features from a sketch developed by a practitioner and associates these to audio features for retrieval of audio units and to synthesis parameter for signal processing. Two mappings between audio and visual features were developed based on findings from previous studies that examined audio and visual feature correlation. One of the mappings is achromatic as the features extracted from the sketch are mainly volumetric and spatial, while the second mapping is chromatic as the features extracted from the sketch are based on color attributes. A number of simple algorithms are discussed that were developed to address three problems: (i) to estimate high level visual feature, (ii) set constrains in the audio corpus to improve the selection algorithm and the exploration of the corpus, and (iii) automatically adjust the weights to improve the efficacy of the selection algorithm in assessing similarity, by optimizing the algorithm in a corpus depended manner.

1. INTRODUCTION

Technological developments in feature extraction, classification, modeling, and data mapping enable us to experience and interact with modal and amodal information in novel ways. These developments have increased our abilities to comprehend and assimilate information and offer many interesting directions for interaction with information, and the physical environment. These developments however have also given rise to questions regarding the design of cross-modal mappings. Are there underlying principles based on which crossmodal associations could be made in objective terms? This research project considers it is of paramount importance to achieve an intuitive mapping to enable interaction with concetenative synthesis for creative purposes (e.g. sound design, electroacoustic composition). The aim of the interface which is presented in this paper is to enable synthesis of sound and the expression of compositional intention by providing perceptually meaningful visual description of sound properties. The intention is to create an objective system for association of visual and aural features drawing on recent findings from studies of perception and cognition. Moreover computational problems specific to feature based synthesis must be addressed in order to enable intuitive control. A number of issues specific to corpus based synthesis can be identified including: (i) partitioning the feature space to avoid empty or unwanted areas of the corpus and improve navigation in the feature space, (ii) controlling the weights to enforce particular feature over others for the selection of audio units, (iii) mapping the distances between target and selected feature vectors to the synthesis parameters in order to modify the selected audio-unit matching as closely as possible the features vector requested by the target. The interface presented in this paper proposes a number of solutions to these issues.

2. MORPHEME INTERFACE

Morpheme is an interface developed using the visual programming language Max/MSP. Morpheme allows to control corpus based concatenative synthesis (see [1]) through the act of sketching on a digital canvas (see figure 1). The implementation of concatenative synthesis that *Morpheme* uses works by segmenting a number of audio files into small units. The units are then analysed, tagged with the analysis data, and stored in a database. Synthesis is accomplished by recombining audio units from the database based on a target feature data stream. Morpheme uses as target the data that derive from the statistical analysis of the sketch's pixel matrix. A number of estimators are deployed to extract high level features from the canvas, and two approaches have been devised in order to improve the exploration of the feature space. Videos demonstrating the interface are available, please see: http://inplayground.wordpress.com/.

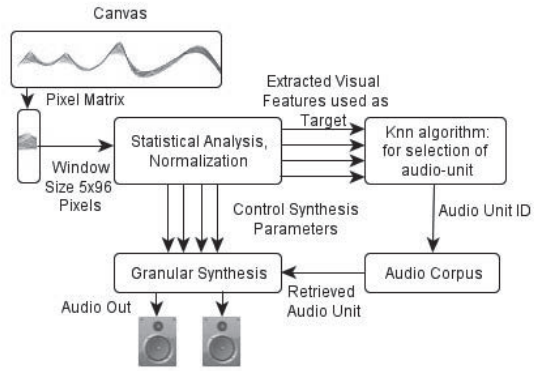

Figure 1. An overview of the architecture of Morpheme

2.1. Parameter Mapping

In the current implementation of *Morpheme,* we can distinguish between three mapping layers. The first layer consists of a mapping between visual and auditory features for the selection of audio units. The second layer consists of mapping the distances between audio and visual features to the synthesis parameters. The third layer is concerned with constraining the target to areas of the audio corpus defined by the user.

2.1.1. Mapping Visual to Audio Features for Selection of Audio Units

Table 1 illustrates the findings from a number of studies that investigated the perceived correlation between auditory and visual features [2], [3], [4], [5]. Some consistency in the feature correlates across the studies can be identified when looking at the table. This consistency is rather encouraging as it suggests that gathering more empirical evidence by deploying diverse methodological approaches could help in the formation of a theoretical framework for the association between visual and aural structures and information. The most common audio-visual correlates according to the empirical findings presented in table 1 were selected to develop two mappings that enable the selection of audio units from the corpus. After several informal trials where different feature set combinations were tested, these two mappings were considered as the most intuitive ones. The distinction between the two mappings is that one is achromatic while the other is chromatic. The first mapping could be considered achromatic in the sense that the visual features extracted from the sketch are estimated based on volumetric and spatial attributes of the sketch (see table 2). The second mapping could be considered as chromatic due to the fact that all of the visual features extracted from the sketch are estimated based on color attributes (see table 3).

Author	Auditory Features	Visual Features
Walker	Loudness	Size
	Pitch	Vertical position
	Timbre	Pattern
	Duration	Horizontal length
Lipscomb et al.	Loudness	Size
	Loudness	Color hue
	Pitch	Vertical position
	Pitch	Color hue
	Timbre	Shape
Giannakis	Loudness	Color saturation
	Pitch	Color brightness
	Dissonance	Texture repetitiveness
	Sharpness	Texture coarseness
	Compactness	Texture granularity
Küssner et al.	Pitch	Vertical position
	Thickness	Loudness
	Time	Horizontal length

Table 1. Highly rated auditory and visual feature pairs.

Audio Features	Visual Features
Spectral flatness	Texture granularity
Pitch	Vertical position
Periodicity	Texture variance
Loudness	Size/Thickness
Duration	Horizontal length

Table 2. Achromatic (i.e volumetric/spatial) mapping between audio and visual features.

Audio Features	Visual Features
Spectral flatness	Color flatness
Spectral Centroid	Color temperature
Periodicity	Color variance
Loudness	Opacity
Duration	Horizontal length

Table 3. Chromatic mapping between audio and visual.

2.1.2. Visual Feature Extraction

The matrix that contains the HSL (i.e. Hue, Saturation, and Lightness) values of the sketch is scanned vertically using a window size of 5 x 96 pixels (see figure 1). During playback, the window slides over the matrix of the canvas, one pixel at a time every 40 milliseconds. A number of statistical analyses are performed in real time on the matrix of the running window. Histogram analysis is performed on the HSL matrix. Morpheme uses the coefficient of variability and kurtosis of the histogram as target features. Besides, *Morpheme* uses three computer vision algorithms to detect the thickness/ size of the painted areas in the window, the centre of the painted area and the texture granularity (see [6], jitter objects cv.jit.mass, cv.jit.centroid, cv.jit.perimeter). Table 4 shows all the features that are extracted from the canvas and explains how they are estimated.

Visual Features	Analysis performed on the 5x96 pixel window to estimate visual features
Size/Thickness	Estimated based on the weighted sum of all ON pixels in a binary image.
Vertical position	Estimated based on the centroid of all ON pixels in a binary image.
Texture granularity	Estimated by dividing the size by the weighted sum of all the edges pixels detected in the window.
Texture variance	Average delta between current and previous window of the edges detected in a binary image.
Colour flatness	HSL histogram flatness (i.e. kurtosis).
Colour variance	HSL histogram coefficient of variance.
Opacity	Alpha mean value of all the pixels.
Colour temperature	Hue mean, occluding background pixels (i.e. white pixels)

Table 4. Visual feature extracted from the canvas.

2.1.3. Mapping the Distances to the Synthesis Parameters

Table 5 shows how the distances between audio features and their respective visual features shown in tables 2 and 3 are associated with the synthesis parameters. The distances are estimated by subtracting the target and selected feature vectors. Some of the decisions regarding the mapping which are presented in table 5, were informed by the initial mapping for the retrieval of audio-units. For example, if the feature of thickness in the visual domain is mapped to loudness, then the distance between target thickness and selected loudness is mapped to control the amplitude parameter of the sound synthesis. However, other audio features such as spectral flatness, periodicity and spectral centroid require more careful consideration as they do not have a direct corresponding synthesis parameter. The decision on how to map the features which do not have a direct correspondence was made in an intuitive manner, by trying different combinations and assessing which correlations are plausible. However to answer these questions in objective terms, empirical work will have to be conducted, to test which correspondences are considered optimal.

Audio features	Synthesis parameters
Spectral flatness	Transposition randomness
Periodicity	Grain size and amplitude randomness
Pitch, Spectral Centroid	Transposition
Loudness	Amplitude

Table 5. Mapping the distances between audio and visual feature vectors to synthesis parameters.

2.1.4. Controlling the Weights for Feature Selection

For the selection of audio units from the corpus, knn (k-nearest neighbor) is used. Knn works by estimating the shortest distance between the feature vector of the target (e.g. visual features extracted from the canvas) and the feature values of the units stored in the database (for more information on the algorithm, see [1], [7]). Because the retrieval of audio units is based on multiple features, often the selected audio-units are not the best match for each individual feature value. Instead, it is the optimal match, taking into consideration all the distances between the target feature vector and the values of the audio-units found in the database. To weigh the selection algorithm in this context means to determine how dependent should the knn algorithm be on a particular feature when estimating the shortest distance between the target feature vector and the feature values of the audio units found in the corpus. A simple method was devised to automatically adjust the weights of each audio feature based on the percentage of dispersion of the audio units for each feature dimension. The weighting algorithm is based on the notion that when a feature value becomes very common between a set of objects, it becomes less salient for establishing links between two or more objects. Consequently it could be argued that feature dimensions that have high

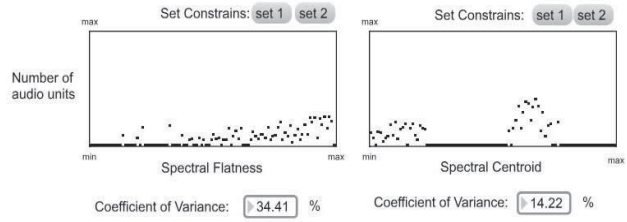

Figure 3. The histograms display the distribution of the audio units in a corpus for two features.

dispersion should be given less weight (i.e. be enforced) than feature dimensions that have low dispersion, as the latter are not distinct enough for assessing feature based similarity. As an example, the histogram of the spectral centroid presented in figure 3 has relatively low dispersion while spectral flatness has higher dispersion; in this case the spectral centroid should be given more weight than spectral flatness. The aim is to weaken feature dimensions when the audio-units stored in the corpus are very similar. As the more similar the audio-units are the less concerned we should be about which audio-unit will be selected. Conversely enforcing the dimensions in which variation of audio material can be found.

This approach helps improve the efficacy of the otherwise distance based algorithm (i.e. knn) in assessing feature-based similarity in a multidimensional context, by optimizing the selection algorithm in a corpus depended manner. To achieve automatic weighting, the coefficient of variation for each feature dimension is estimated based on its histogram. The coefficient of variation is the ratio of standard deviation to the mean. It provides an estimate of the variability of the audio-units in the corpus which help compare the audio features that have different mean value. The percentage of dispersion given by the coefficient of variation is used to determine how dependent should the selection algorithm be when assessing the distances.

2.1.5. Constraining the Selection Algorithm

Constraining the selection algorithm to the areas of the corpus where audio units have formed clusters can improve the navigation of the feature space. A common phenomenon when an audio corpus consists of a relatively small number of audio units is that the distribution of the audio descriptions in the feature space is concentrated forming dense clusters in some areas while the rest of the feature space is relatively empty (see figure 4). One problem that can be observed is that the target features values requested might be well outside the main clusters of the feature space and the target might not match any of the audio units in the corpus. As mentioned earlier in such case, the knn algorithm will select the nearest unit that can be found in the feature space. On the one hand this is very useful as a unit can be selected even if the target does not exactly correspond to any of the descriptions of the audio units found in the corpus. On the other hand if the target feature vectors in a series of queries are well outside the main clusters, the selection algorithm will stay in the periphery of the clusters and will not access the clusters (i.e. figure 4 shows what is referred to as periphery). This results in the following problems: (i) although the corpus might consist of many audio-units, it might be

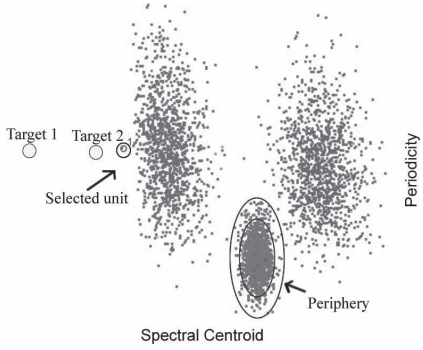

Figure 4. A two dimensional plots shows the distribution of the audio units in a corpus.

difficult to access the clusters, and (ii) two very different target queries (i.e. two brush strokes in the context of *Morpheme*), which both request feature vectors that are well outside the main clusters of the corpus, might retrieve the very same audio-unit. Figure 4 demonstrates the problem; *Target 1* and *Target 2* are relatively distant in the feature space, however the same audio unit is selected as it is the nearest. This in a sense makes difficult to explore the feature space and it can create ambiguities for the user regarding the association between target and selected feature.

A simple method has been devised to address these issues and improve the navigation of the feature space. To achieve this, a histogram analysis is performed across each of the features dimensions which are used for retrieval, like in figure 5. The histogram shows the distribution of all the audio units of a particular feature. Currently *Morpheme* allows to set up two constraints in order to avoid empty or unwanted areas of the corpus (e.g. empty, silent or undesired audio units). This is accomplished by setting minimum and maximum bounds using a pointing device directly on the histogram's graphical representation (see figure 5). The minimum and maximum bounds are then used to scale the target features and map them to the areas of the corpus defined by the constraints which were set by the user. So any target query that requests feature vectors from the *Area out of bound 1* will be scaled to a corresponding value from *Constrain area to bounds 1,* while queries that request units from *Area out of bound 2* are mapped to *Constrain area to bounds 2.*

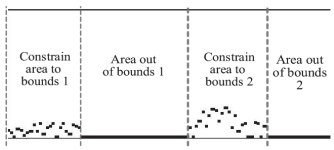

Figure 5. A two dimensional plots display the distribution of the audio units and the constraints.

3. DISCUSSION & FUTURE WORK

Two audio to visual mappings were proposed drawing on previous empirical work on audio-visual feature correspondence. Evaluation will be necessary to further assess the audio and visual mappings. Studies will include testing similarity between auditory and visual features in different contexts, such as testing individual feature correlations, correlations in multidimensional feature sets, and testing how perceived correspondence might be affected if the corpus content is different. The constraints approach presented in this paper could be improved by automatically detecting cluster and by remapping the target query when it requests feature values that are outside the bounds to the nearest constraint area. The corpus depended approach for adjusting the weights based on feature dispersion, could be improved by taking into consideration other statistical measurements such as the number of clusters, and their densities for each feature dimension. Much can still be done to improve the navigation of the feature space and address the issues discussed in this paper. For example providing finer control of the brush parameters by defining high level attributes (i.e. texture granularity/ repetitiveness/ coarseness, color variance/flatness) will be required to give practitioners to more precise control over the exploration of the audio corpus. Further the visual feature estimators presented in this article approximate visual attributes for which we do not currently have a commonly accepted model for descriptive purposes. In the future, further elaboration of estimates in conjunction with empirical testing will be necessary. More specifically there is need to construct more elaborate models for the interpretation of the visual descriptors that derive from the sketch and create more complex relationships between the descriptors data, with the aim to extract higher level visual features.

4. ACKNOWLEDGMENT

The author would like to acknowledge the Institute for Informatics & Digital Innovation for their financial support and Diemo Schwarz and the IMTR team for sharing CataRT.

5. REFERENCES

[1] D. Schwarz, G. Beller, B. Verbrugghe. S. Britton, Real-Time Corpus Based Concatenative Synthesis with CataRT. *In proceedings of DAFx*, 2006.

[2] R. Walker, The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception and Psychophysics,* Vol 42(5), pp. 491–502, 1987.

[3] S. D. Lipscomb, E. M. Kim, Perceived match between visual parameters and auditory correlates: an experimental multimedia investigation. ICMPC 2004.

[4] K. Giannakis, A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound Journal,* vol. 11, no. 3, pp. 297–307, 2006.

[5] M. B. Küssner, H. M. Prior., N. E. Gold.,D. Leech-Wilkinson. Getting the shapes "right" at the expense of creativity? *ICMPC & ESCOM, 2012*

[6] J. M. Palletier, CV.Jit: http://jmpelletier.com/cvjit/.

[7] D. Schwarz, Data-Driven Concatenative Sound Synthesis. PhD thesis, Universite Paris 6, 2004.